Mutation rates in mammalian genomes

Sudhir Kumar*† and Sankar Subramanian*

Department of Biology, Arizona State University, Tempe, AZ 85287-1501

Knowledge of the rate of point mutation is of fundamental importance, because mutations are a vital source of genetic novelty and a significant cause of human diseases. Currently, mutation rate is thought to vary many fold among genes within a genome and among lineages in mammals. We have conducted a computational analysis of 5,669 genes (17,208 sequences) from species representing major groups of placental mammals to characterize the extent of mutation rate differences among genes in a genome and among diverse mammalian lineages. We find that mutation rate is approximately constant per year and largely similar among genes. Similarity of mutation rates among lineages with vastly different generation lengths and physiological attributes points to a much greater contribution of replication-independent mutational processes to the overall mutation rate. Our results suggest that the average mammalian genome mutation rate is 2.2×10^{-9} per base pair per year, which provides further opportunities for estimating species and population divergence times by using molecular clocks.

neutral evolution | substitution pattern | disparity index | generation length | molecular clock

Rates of point mutation can be determined indirectly by estimating the rate at which the neutral substitutions accumulate in protein-coding genes (1). Synonymous substitutions in protein-coding genes generally are free from natural selection and are used frequently for inferring neutral substitution rates (1, 2). In particular, the fourfold-degenerate sites are expected to harbor only the neutral substitutions, because all mutations at these sites are synonymous at the amino acid sequence level. By using estimates of evolutionary distances based on neutral substitutions, many studies have examined the null hypotheses of uniformity of neutral mutation rates among genes within a genome and among mammalian lineages and have come to conflicting conclusions (2–9). For example, significant differences in mutation rates among mammalian lineages reported over the last two decades led to the proposal of the generation-time effect hypothesis (10–13). However, Easteal *et al.* (14) have argued that previous results of substantial differences among lineages observed may have been caused by the use of incorrect fossil dates or inappropriate outgroups. Similarly, there is significant controversy regarding differences in mutation rate among genes within a genome (8, 9, 15) and over 10-fold differences in the estimates of the mutation rates among studies ($1.1–12.4 \times 10^{-9}$ substitutions per site per year; refs. 3, 11, 12, and 16–19).

One common feature of many of these studies is that they have either analyzed a small number of genes or only a few species. Analysis of a large sample of genes from a genome and diverse phylogenetic lineages is the key to testing the null hypothesis of equal mutation rates within and among genomes. A large number of genes is necessary, because only a fraction ($\approx 15\%$) of codon positions in a sequence are fourfold-degenerate (see Fig. 1 legend) and we need to sample genomic regions extensively. Furthermore, mutation rate information from many inter- as well as intraordinal mammalian species pairs is necessary to test whether the observed differences, if any, among mammalian orders are likely to be tied significantly to differences in generation times and physiological attributes among groups. There-

fore, we have assembled a data set of 17,208 protein-coding DNA sequences belonging to 5,669 different nuclear genes from a total of 326 placental mammalian species to characterize the extent of difference in mutation rates among genes in a genome and among lineages.

Materials and Methods

Data Mining and Assembly. Phylogenetic trees of 8,627 gene families in the HOVERGEN database (20) release 36 were constructed from amino acid sequence alignments by using the neighbor-joining method in MEGA2 (21). The cDNA sequence alignments for orthologous sequence sets then were generated using amino acid sequence alignments as guides. Neighbor-joining trees were scanned automatically followed by manual inspection to identify orthologous sequence sets. We enforced strict orthology definitions by considering sequences to be orthologous only if no gene duplication events were detected since their divergence from the most recent common ancestor. All gene families containing fewer than three sequences were excluded, which produced a set of 3,132 gene families. There were a total of 326 species, with 113 species represented by >2 genes. The final data set assembled 17,208 protein-coding DNA sequences belonging to 5,669 different nuclear genes from a total of 326 placental mammalian species available in the databanks. The number of sequences available from different species and groups varied extensively: primates, 5,618 sequences; sciurognath rodents, 8,142 sequences; artiodactyls, 2,042 sequences; lagomorphs, 573 sequences; carnivores, 483 sequences; hystricognath rodents, 205 sequence; and perissodactyls, 145 sequences.

All computations were done by using only the fourfold-degenerate sites for sequence pairs. We took a stringent approach in identifying fourfold-degenerate sites by selecting only those sites that have remained fourfold-degenerate throughout the evolutionary history of the pair of species compared. This task was accomplished by designating a site as fourfold-degenerate only if it was so in both the sequences compared. Fig. 1 shows the distribution of the number of fourfold-degenerate sites in 3,722 genes in the human-mouse comparison.

Estimation of Evolutionary Distance. Evolutionary divergence (d_4) between sequences at fourfold-degenerate sites was estimated by using the Tamura–Nei method (22) to correct for multiple hits by accounting for transition/transversion rate and base-frequency biases. For a given species pair, multigene evolutionary distance was computed by taking the average of evolutionary distance over all genes. For two groups of species, the evolutionary distance was estimated by first computing average gene distances between species belonging to the two groups and then taking an average of these distances over all genes.

EVOLUTION
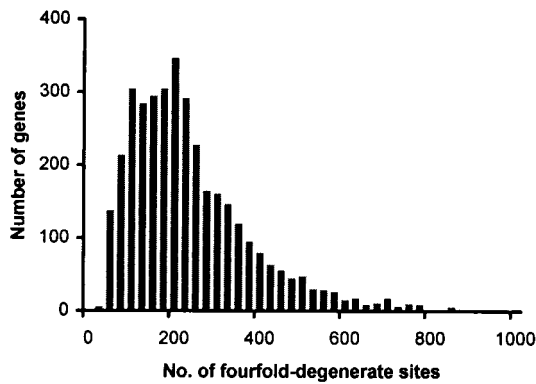
Fig. 1. Distribution of the number of fourfold-degenerate sites in the human-mouse orthologous genes. The average number of fourfold-degenerate sites from 3,722 genes of human-mouse comparison is 245, which is 15.7% of the average cDNA length (1,568).

Estimation of Expected Variance for a Multigene Distribution. The expected amount of variance in the distribution of multigene distances for a given pair of species is the sum of the estimation variation (V_e) contributed by the use of distance methods to correct for multiple hits and the variance contributed by the stochastic nature of the evolutionary process (V_s). For a set of N independent genes, $V_e = \Sigma_i(V_{e,i})/N$, where $V_{e,i}$ is the estimation variance for gene i. It is computed by using Tamura–Nei's variance formula (22), V_{TN}, as $V_{e,i} = V_{TN} - d_i/L_i$, where L_i is the number of sites, and d_i is the Tamura–Nei distance. Under the null hypothesis of equal mutation rate per site (μ), V_s is obtained by considering a Poisson process governing the arrival of mutations at a finite number of fourfold-degenerate sites in a given gene. For N genes, it is given by $V_s = \Sigma_i (\mu/L_i)/N$.

Determination of Physical Location of Genes in the Human Genome Map. Currently the human genome map consists of relative positions of a large number of contigs on each chromosome. The gene content of each contig was obtained from the NCBI ftp site (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/). We first mapped the GenBank accession numbers of human genes in our data set to their corresponding unique LocusIDs by using the LocusLink public resource (http://ncbi.nlm.nih.gov/locuslink) and then constructed a complete map for all human sequences included in this study. Chromosomal locations of mouse genes also were obtained from LocusLink. For the analysis of gene proximity and mutation rate, the physical distance between a gene pair on human chromosome was estimated by subtracting the ending nucleotide position of the first gene from the starting nucleotide position of the second gene. Only adjacent genes on human chromosome belonging to the same conserved synteny in the mouse genome were used in the gene pairs, and no gene was used more than once. Conserved syntenies were determined by using the contiguous sets of autosomal markers method (23).

Results and Discussion

Homogeneity of Substitution Patterns Between Lineages. Although the fourfold-degenerate sites are expected to accumulate only synonymous substitutions, the evolutionary distances estimated by using these sites are useful in estimating the underlying mutation rate only if the nucleotide substitutions have accrued with the same substitution pattern in the two species compared. That is, the homologous sites in the two sequences compared *in a given gene* must have evolved with the same instantaneous substitution matrices. Substitution patterns in a given gene may shift in one lineage as compared with its orthologous counterpart for a number of reasons including chromosomal rearrangements

(23), gene transfer (24), or centromere movement (e.g., mouse genome). In these cases, substitution patterns in genes may be affected to fix mutations that make the base composition of the gene to be more similar to its chromosomal location [amelioration effect (24)], and this will be more pronounced at the sites that are selectively neutral. Therefore, the substitution rate at neutral sites in those genes will be higher than the actual mutation rate (25), rendering such genes unsuitable for inferring mutation rates.

Therefore, we conducted the disparity index test for each pair of orthologous sequences (26, 27) to identify genes in which fourfold-degenerate sites are not evolving with homogeneous substitution patterns among the lineages compared. The disparity index test directly examines the null hypothesis of homogeneity of the evolutionary pattern between two lineages by testing whether the observed difference in nucleotide frequencies between sequences is more than that expected by chance alone, given the number of differences observed between sequences. It does not require the knowledge of the actual pattern of substitution, evolutionary relationships among species, or equality of substitution rates among lineages (26, 27).

The disparity index test revealed that the fourfold-degenerate sites in a large number of genes have not evolved homogeneously in inter- as well as intraordinal comparisons (Fig. 2a). For instance, sequences of the same gene in human and mouse are evolving with significantly different evolutionary substitution patterns in 1,703 of 3,722 comparisons (46%). The red closed circles in Fig. 2b correspond to genes that were rejected by the disparity index test when the expected and observed difference in GC content was tested. These genes clearly show much higher observed GC content difference than that expected by chance alone (the expected distribution is depicted by green open circles). On the contrary, GC content differences between human and mouse for the genes passing the disparity index test (black closed triangles) show a distribution that overlaps with the expected distribution (Fig. 2b). It is apparent that mutations in the fourfold-degenerate sites are fixed with different patterns of substitution in different genes depending on the chromosomal context (e.g., isochore structure) in the genome (28). However, the observed differences in G + C content in fourfold-degenerate sites among genes may be an indication of differences in actual patterns of substitution or mutation among genes.

Therefore, synonymous substitutions in a large number of genes are not suitable to use for inferring mutation rates. In fact, the inclusion of genes (sequence pairs) evolving under heterogeneous evolutionary patterns would produce distance estimates that are higher than that expected for the genes evolving with homogenous substitution patterns. Because different lineage comparisons show this heterogeneity to different extents (Fig. 2c), estimates will be biased to different extents, which is likely to lead to erroneous conclusions regarding large mutation-rate differences among species (25). Both these problems are clearly evident in Fig. 2c, which shows that the difference in evolutionary distances at fourfold-degenerate sites among the genes, which passed or failed the disparity index test, is as large as 46% (cow-pig comparison) and differs multifold among different species pairs. For this reason, all genes showing pattern heterogeneity in fourfold-degenerate sites should be and were removed from any further analyses. This removal reduces the number of genes considerably (Fig. 2a), but still the numbers of the fourfold-degenerate sites analyzed were quite large (682–543,962; see Fig. 5 legend).

Distribution of Evolutionary Rates Among Genes in Human and Mouse. Fig. 3a shows the distribution of the neutral distance estimated by using the fourfold-degenerate sites (d_4) from 2,019 human-mouse genes. This distribution of neutral distance is equivalent to the distribution of evolutionary rates, because the time of
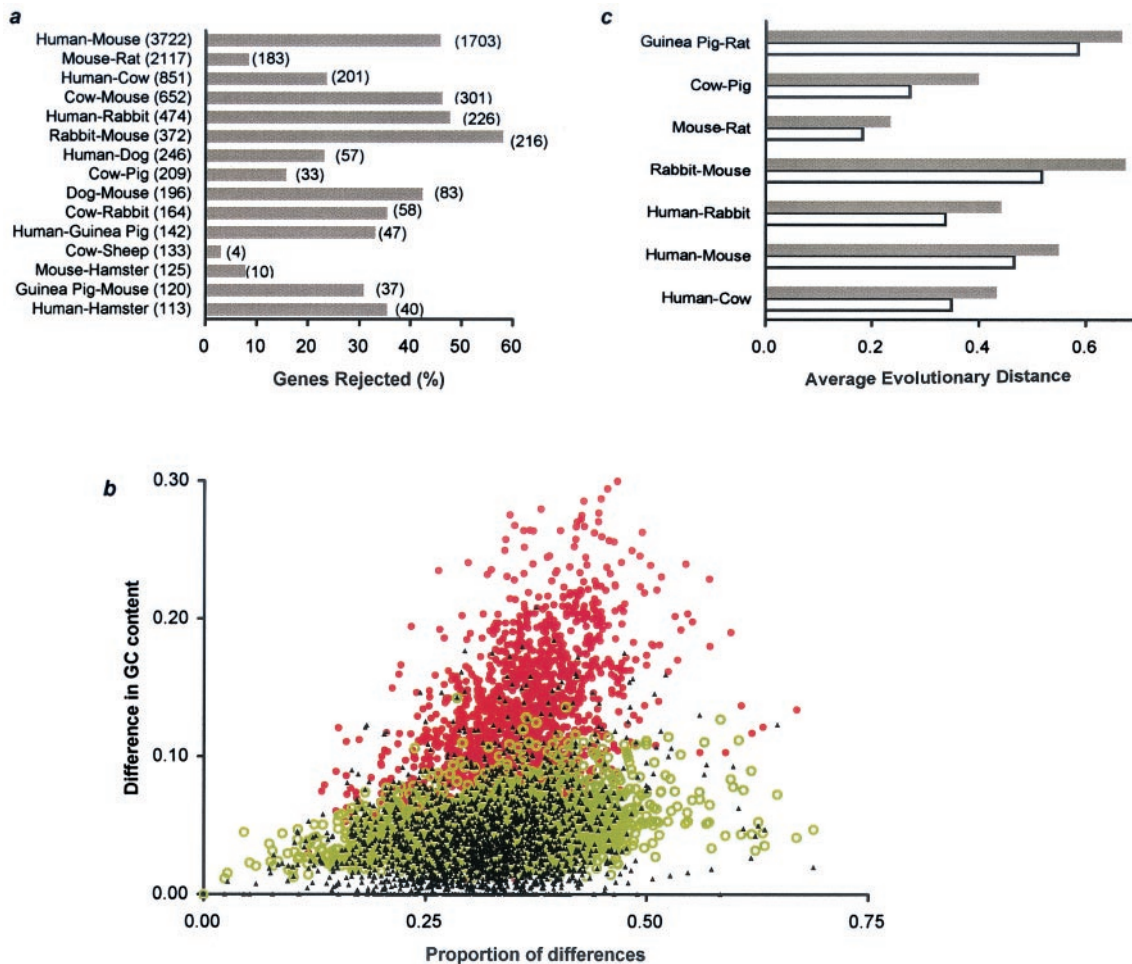
Fig. 2. (*a*) Proportion of genes rejected for 15 mammalian pairs using the disparity index test in testing the null hypothesis of homogeneity of substitution patterns in fourfold-degenerate sites at a 5% significance level (26). The total number of genes examined is shown in parentheses next to the name of the species pair; the number of genes rejected is shown next to the corresponding bar. (*b*) Scatter plot showing the relationship of ΔGC_4 (GC content difference at fourfold-degenerate sites in orthologous sequences) with p_4 (proportion of fourfold-degenerate sites with A/T $\leftrightarrow$ G/C difference) for genes passing (black) and failing (red) the disparity index test. Green markers show the expected relationship of ΔGC_4 calculated from sequence length and p_4 for each gene by using equation 10 in ref. 26. (*c*) Overestimation of evolutionary distance caused by genes evolving with heterogeneous patterns of evolution for seven representative mammalian pairs. Open bars show the estimate of average distance from all genes passing the homogeneity test (26, 27). Dark bars show the average distance for the rejected genes (heterogeneous).

divergence is the same for each pair of orthologous sequences and thus is a constant factor in each comparison. This distribution is largely symmetrical and bell-shaped and has considerable dispersion (mean = 0.466, variance = 0.035). We find that genes with small sequence lengths mostly contribute to this variation, because long genes show distance estimates close to the overall average as compared with the short genes (Fig. 3*b*). Importantly, the averages for short and long genes are almost identical to the overall average. In fact, the expected variance of 0.029 (0.025 and 0.004 for estimation and stochastic variances, respectively; see *Materials and Methods*) of the multigene distribution under the null hypothesis of uniform mutation rate among genes is close to the observed variance (0.035). A normal curve drawn with expected variance around the mean appears to enclose the observed distribution well (Fig. 3*a*). Therefore, the observed variation in neutral distances among genes can arise even if the mutation rates are the same among genes.

We directly tested the null hypothesis of mutation rate uniformity among genes by examining the relationship of gene proximity and evolutionary distance by using 1,901 genes, for which the exact location on human was available and the

information about their relative position on a segment of conserved synteny with mouse was known. If the genes located closely evolve with similar rates as reported (8, 9), the difference in neutral evolutionary distance (Δd_4) of closely located gene pairs would be smaller than that for distantly located gene pairs. To test this prediction, we computed the average Δd_4 of gene pairs located in 1–5 million base-pair (Mbp) distance with an interval of 0.5 Mbp. Fig. 4*a* shows that the average Δd_4 for closely located genes is not smaller than those located further apart. For example, the average Δd_4 for genes located less than 0.5 Mbp apart and less than 5.0 Mbp (but more than 4.5 Mbp) apart is very similar (0.17 and 0.15, respectively; Fig. 4*a*). In fact there is no significant correlation between the gene proximity and evolutionary rate plotted in Fig. 4*a*. This holds true in the analyses using the genes located even closer (Fig. 4*b*). The average evolutionary distances in different conserved segments are close to the overall average distance estimated using all the homogeneously evolving human-mouse orthologous genes (Fig. 4*c*). However, apart from a few outliers, chromosome 19 does show significantly higher evolutionary distances compared with other chromosomes. As expected, the average distance for the human
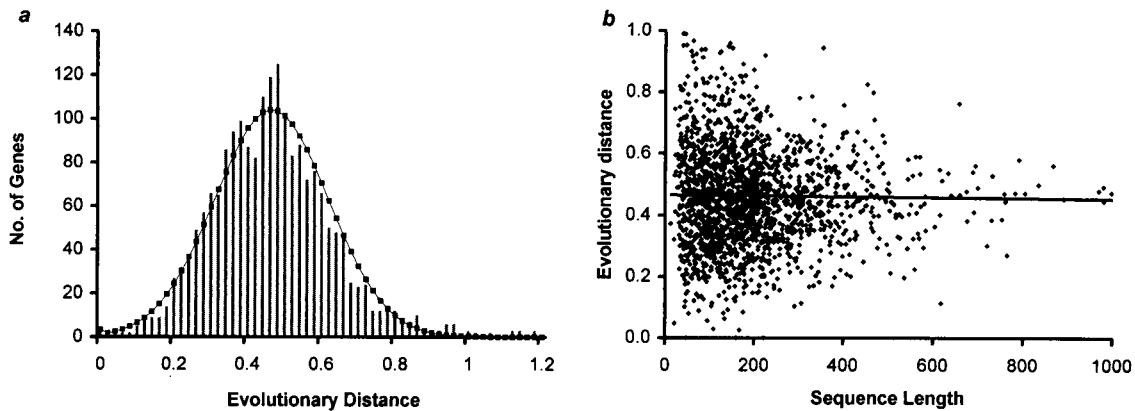
Fig. 3. (*a*) A histogram of the neutral evolutionary distances estimated by using fourfold-degenerate sites of 2,019 human and mouse genes (mean = 0.466 and variance = 0.035; two outliers were removed from each tail). The normal distribution curve is drawn by using the observed mean and the expected variance under the null hypothesis of uniform neutral mutation rate among genes. (*b*) The distribution of evolutionary distances estimated from genes of varying sequence lengths.

X chromosome was lower than the average mutation rate (0.37 as compared with 0.47).

Mutation Rates Among Mammalian Lineages. We examined the temporal constancy of neutral evolutionary rate among diverse mammalian lineages. The divergence time and evolutionary distance for 43 mammalian pairs clearly show a linear relationship (correlation coefficient = 0.97) with the regression analysis indicating that the mammalian genomes accumulate mutations at an average rate of 2.22×10^{-9} substitutions per site per year (Fig. 5*a*). We also estimated the mutation rate by using only

fossil-based divergence times for intraordinal splits and 90 million years for superordinal divergences (29–31). This yields an upper bound of 2.61×10^{-9} substitutions per site per year (Fig. 5*b*). We did not include the fossil-based estimates for the divergence within scuirognath rodents because of their highly controversial nature (14, 32–36). However, inclusion of 10–16 million years divergence for among murid rodent families did not change the slope of the regression line for the fossil-based estimates.

To obtain an assessment independent of the divergence times, we examined the similarity of rates among lineages by conducting
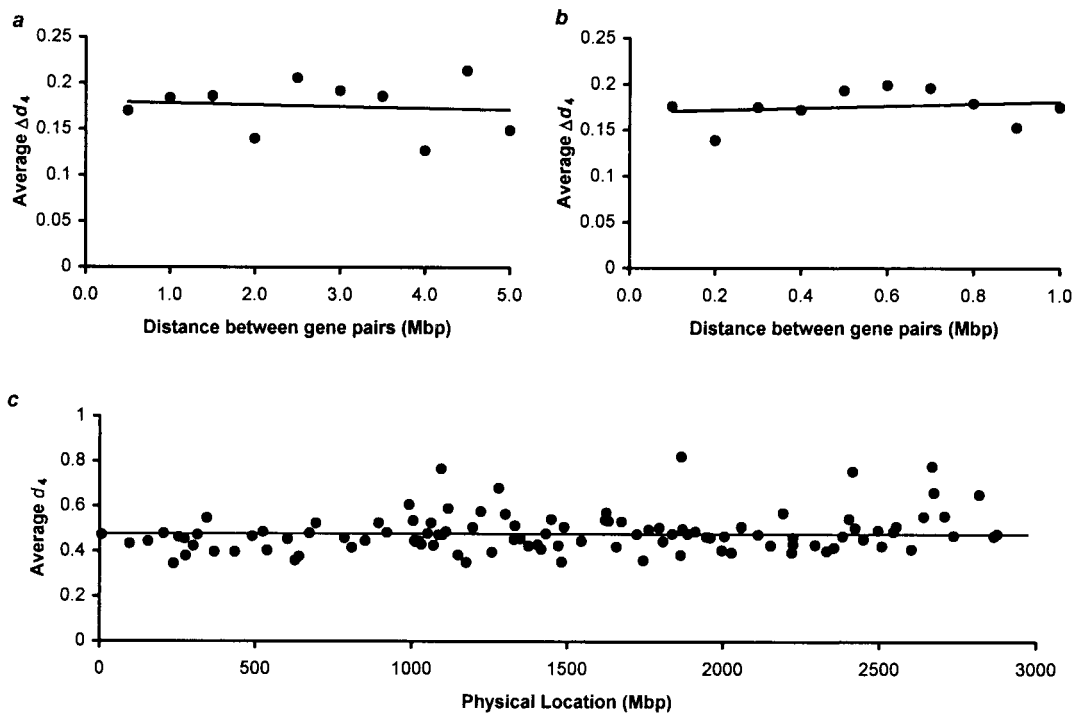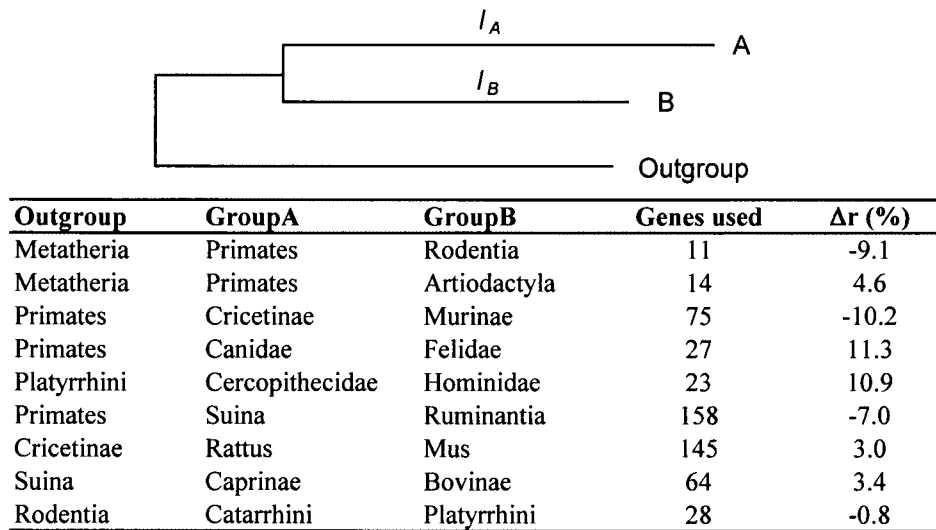

Fig. 4. Average differences in neutral evolutionary distances using human-mouse comparison (Δd_4) were plotted for gene pairs located within 0.5 (*a*) and 0.1 Mbp (*b*). (*c*) Average neutral evolutionary distances (d_4) from the genes located within a human-mouse conserved segment were plotted corresponding to their position on the human chromosomes. The *x* axis refers to the starting location of the conserved segments in a super chromosome where all 22 autosomes were concatenated in order from head to tail. A total of 108 conserved segments, each containing a minimum of five genes, were used. The horizontal line indicates the average d_4 computed by using all 1,901 genes.

Kumar and Subramanian
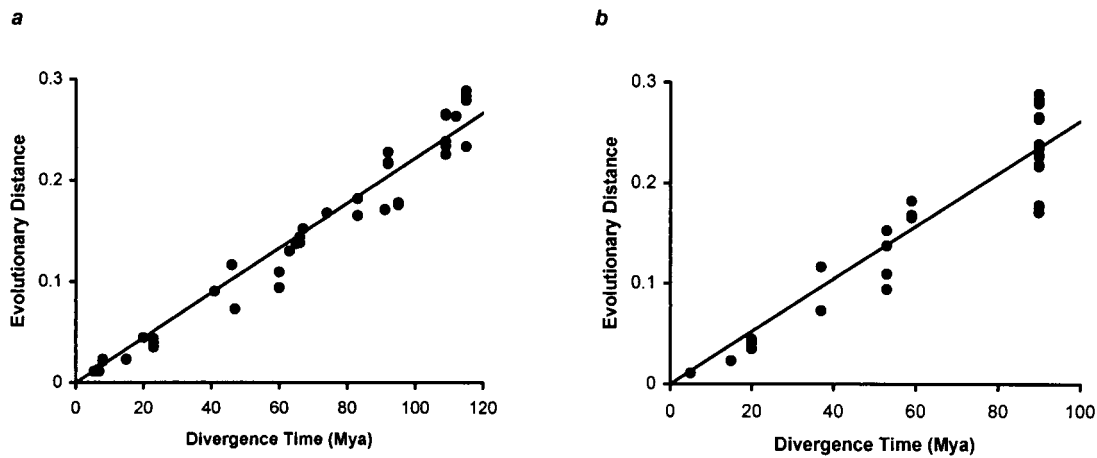
Fig. 5. Accumulation of neutral substitutions over time (million years) in diverse mammalian species. The average distance per lineage is plotted for inter- and intraordinal comparisons by using, respectievely, molecular data-based (*a*) and fossil-based (*b*) time estimates reported previously. Linear regression produced a rate of 2.22×10^{-9} substitutions per site per year million years ago ($r = 0.97$, $n = 43$) for the molecular data and 2.61×10^{-9} substitutions per site per year ($r = 0.92$, $n = 33$) for the fossil data. The divergence time estimates (million years ago, Mya) from molecular data (30, 35, 40–43)/fossil records (44–46), and the number of fourfold sites/genes used were: human-chimpanzee (5.5/5.5, 6,302/33), human-gorilla (7/–, 3,605/23), human-orangutan (8/–, 4,329/30), chimpanzee-gorilla (7/–, 1,674/13), chimpanzee-orangutan (8/–, 1,497/10), gorilla-orangutan (8/–, 1,936/15), cercopithecidae-hylobatidae (23/20, 1,290/10), hominidae-hylobatidae (15/–, 2,321/19), cercopithecidae-hominidae (23/20, 38,128/231), catarrhini-platyrrhini (47/37, 8,194/54), cetacea-ruminantia (60/53, 955/8), cetacea-suina (60/53, 682/10), ruminantia-tylopoda (67/53, 736/5), ruminantia-suina (65/53, 30,063/204), bovinae-caprinae (20/20, 24,474/158), bovoidea-cervoidea (23/20, 1,765/16), canidae-felidae (46/37, 5,651/41), gerbillinae-cricetinae (66/–, 1,172/11), gerbillinae-murinae (66/–, 2,605/24), murinae-cricetinae (66/–, 52,043/264), mus-rattus (41/–, 486,306/2,139), primates-carnivora (95/–, 49,267/266), primates-cetartiodactyla (95/–, 179,032/985), primates-perissodactyla (95/–, 11,593/75), primates-lagomorpha (91/–, 54,670/260), primates-hystricognath rodents (109/–, 18,292/110), primates-sciurognath rodents (115/–, 543,962/2,708), cetartiodactyla-carnivora (83/–, 24,086/160), cetartiodactyla-perissodactyla (83/–, 9,188/75), cetartiodactyla-lagomorpha (92/–, 30,465/175), cetartiodactyla-hystricognath rodents (109/–, 10,227/72), cetartiodactyla-sciurognath rodents (115/–, 119,345/756), carnivora-perissodactyla (74/–, 5,364/44), carnivora-lagomorpha (92/–, 9,498/58), carnivora-hystricognath rodents (109/–, 4,986/31), carnivora-sciurognath rodents (115/–, 28,644/191), perissodactyla-lagomorpha (92/–, 4,131/34), perissodactyla-hystricognath rodents (109/–, 1,779/18), perissodactyla-sciurognath rodents (115/–, 7,843/60), hystricognath rodents-lagomorpha (109/–, 5,447/36), hystricognath-sciurognath rodents (115/–, 19,930/123), sciurognath rodents-lagomorpha (112/–, 36,672/219). In the above, a dash indicates the absence of information.

relative rate tests using an outgroup species (Fig. 6). For primates and rodents (using marsupials as an outgroup), the average rate difference in multigene analysis was ≈9%. Interestingly, a similar magnitude of difference was found even in intraordinal comparisons. For instance, the average rate difference between canidae and felidae or between murinae and cricetinae using primates as outgroup is 10–11%. Therefore, small rate differences seem to exist among lineages, and clearly there are no

Outgroup	GroupA	GroupB	Genes used	Δr (%)
Metatheria	Primates	Rodentia	11	-9.1
Metatheria	Primates	Artiodactyla	14	4.6
Primates	Cricetinae	Murinae	75	-10.2
Primates	Canidae	Felidae	27	11.3
Platyrrhini	Cercopithecidae	Hominidae	23	10.9
Primates	Suina	Ruminantia	158	-7.0
Cricetinae	Rattus	Mus	145	3.0
Suina	Caprinae	Bovinae	64	3.4
Rodentia	Catarrhini	Platyrrhini	28	-0.8

Fig. 6. Relative evolutionary rate among different mammalian lineages. The relative evolutionary rate differences (Δr) in a gene for two lineages (say, *A* and *B*) were computed as $\Delta r = [l_A - \frac{1}{2}(l_A + l_B)]/\frac{1}{2}(l_A + l_B)$, where l_A and l_B are the least-squares estimates of the branch lengths leading to the two lineages using the average distances between species/groups *A* and *B* and their distances with the outgroup(s). Δr was computed for each gene, and the mean value taken over all genes for groups *A* and *B*. Negative values indicate that group *B* is evolving faster than group *A*. Only the genes passing the disparity index test were used in this analysis.

systematic relationships between the evolutionary rate and generation length. This means that the generation length and physiological differences among diverse groups do not influence the neutral substitution rates significantly, and the evolutionary time is the principal factor dictating the accumulation of neutral mutations.

A much larger rate difference between mice and humans has been reported (3, 13). Such results seem to be caused by the inclusion of genes evolving with heterogeneous substitution patterns, because we find that these genes show a much larger relative rate difference (34%) between primates and rodents than those from genes that pass the disparity index test for homogeneity of substitution patterns (9%). Therefore, fixation of mutations under heterogeneous substitution pattern in the orthologous sequences rather than the difference in mutation rates is likely to be the cause for the results reported previously.

The absence of significant correlation between the mutation rate and generation time is likely to prompt a reassessment of the generation time effect hypothesis, in which errors in DNA replication in germ-line cells is considered to be the major source of mutation (11, 13). This hypothesis predicts a higher mutation rate in species with shorter generation time (e.g., mice as compared with humans). This is because, as is thought currently, species with the shorter generation length will undergo more germ-line cell divisions per year and thus accumulate a larger number of replication errors in unit time, which would lead to a larger mutation rate per unit time in those species. However, the relationship of the number of germ-line cell divisions and the mutation rate clearly is not linear as is indicated in the difference in the ratio of the male/female mutation rate in primates and rodents. Primates show a male/female mutation-rate ratio (37, 38) that is almost the same as seen in rodents even when the ratio of the number of germ-cell divisions in males and females is almost 3-fold higher in humans as compared with mice. Therefore, our results suggest that replication-independent mutational processes (e.g., DNA methylation, recombination, and repair mechanism) may play a greater role as a source of mutation than that anticipated earlier (4, 38, 39).

In conclusion, our results argue against the widely held notion about large differences in mutation rates among genes in a genome and among major mammalian lineages. This approximate similarity of mutation rates among genes and among lineages is likely to be important for estimating divergence time for closely related species, testing for selection by comparative sequence analysis, inferring coalescent times, and understanding the mutational processes that govern evolution of mammalian genomes.

1. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
2. Nei, M. & Kumar, S. (2000) *Molecular Evolution and Phylogenetics* (Oxford Univ. Press, New York).
3. Li, W.-H. & Tanimura, M. (1987) *Nature (London)* **326,** 93–96.
4. Britten, R. J. (1986) *Science* **231,** 1393–1398.
5. Wolfe, K. H., Sharp, P. M. & Li, W.-H. (1989) *Nature (London)* **337,** 283–285.
6. Easteal, S. & Collet, C. (1994) *Mol. Biol. Evol.* **11,** 643–647.
7. Mouchiroud, D., Gautier, C. & Bernardi, G. (1995) *J. Mol. Evol.* **40,** 107–113.
8. Matassi, G., Sharp, P. M. & Gautier, C. (1999) *Curr. Biol.* **9,** 786–791.
9. Williams, E. J. & Hurst, L. D. (2000) *Nature (London)* **407,** 900–903.
10. Laird, C. D., McConaughy, B. L. & McCarthy, B. J. (1969) *Nature (London)* **224,** 149–154.
11. Wu, C.-I. & Li, W.-H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1741–1745.
12. Li, W.-H., Tanimura, M. & Sharp, P. M. (1987) *J. Mol. Evol.* **25,** 330–342.
13. Li, W.-H., Ellsworth, D. L., Krushkal, J., Chang, B. H.-J. & Hewett-Emmett, D. (1996) *Mol. Phylogenet. Evol.* **5,** 182–187.
14. Easteal, S., Collet, C. & Betty, D. (1995) *The Mammalian Molecular Clock* (Landes, Austin, TX).
15. Wolfe, K. H. & Sharp, P. M. (1993) *J. Mol. Evol.* **37,** 441–456.
16. Bulmer, M., Wolfe, K. H. & Sharp, P. M. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 5974–5978.
17. Makalowski, W. & Boguski, M. S. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9407–9412.
18. Nachman, M. W. & Crowell, S. L. (2000) *Genetics* **156,** 297–304.
19. Keightley, P. D. & Eyre-Walker, A. (2000) *Science* **290,** 331–333.
20. Duret, L., Mouchiroud, D. & Gouy, M. (1994) *Nucleic Acids Res.* **22,** 2360–2365.
21. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001) *Bioinformatics,* in press.
22. Tamura, K. & Nei, M. (1993) *Mol. Biol. Evol.* **10,** 512–526.
23. Kumar, S., Gadagkar, S. R., Filipski, A. & Gu, X. (2001) *Genetics* **157,** 1387–1395.
24. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44,** 383–397.
25. Saccone, C., Pesole, G. & Preparata, G. (1989) *J. Mol. Evol.* **29,** 407–411.
26. Kumar, S. & Gadagkar, S. R. (2001) *Genetics* **158,** 1321–1327.
27. Kumar, S. & Gadagkar, S. R. (2001) *Genetics* **159,** 913–914.
28. Sueoka, N. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2653–2657.
29. Archibald, J. D. (1996) *Science* **272,** 1150–1153.
30. Hedges, S. B. & Kumar, S. (1999) *Science* **285,** 2031a.
31. Archibald, J. D., Averianov, A. O. & Ekdale, E. G. (2001) *Nature (London)* **414,** 62–65.
32. Jaeger, J. J., Tong, H. & Denys, C. (1986) *C. R. Acad. Sci.* **302,** 917–922.
33. O'hUigin, C. & Li, W.-H. (1992) *J. Mol. Evol.* **35,** 377–384.
34. Janke, A., Feldmaierfuchs, G., Thomas, W. K., Vonhaeseler, A. & Paabo, S. (1994) *Genetics* **137,** 243–256.
35. Kumar, S. & Hedges, B. (1998) *Nature (London)* **392,** 917–919.
36. Nei, M., Xu, P. & Glazko, G. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2497–2502. (First Published February 20, 2001; 10.1073/pnas.051611498)
37. Bohossian, H. B., Skaletsky, H. & Page, D. C. (2000) *Nature (London)* **406,** 622–625.
38. Huttley, G. A., Jakobsen, I. B., Wilson, S. R. & Easteal, S. (2000) *Mol. Biol. Evol.* **17,** 929–937.
39. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148,** 1667–1686.
40. Arnason, U., Gullberg, A., Gretarsdottir, S., Ursing, B. & Janke, A. (2000) *J. Mol. Evol.* **50,** 569–578.
41. Janke, A., Xu, X. & Arnason, U. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 1276–1281.
42. Huchon, D., Catzeflis, F. M. & Douzery, E. J. (2000) *Proc. R. Soc. London Ser. B* **267,** 393–402.
43. Mouchaty, S. K., Catzeflis, F., Janke, A. & Arnason, U. (2001) *Mol. Phylogenet. Evol.* **18,** 127–135.
44. Benton, M. J. (1993) *The Fossil Record 2* (Chapman and Hall, London).
45. Carroll, R. L. (1988) *Vertebrate Paleontology and Evolution* (Freeman, New York).
46. Haile-Selassie, Y. (2001) *Nature (London)* **412,** 178–181.